Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Journal of Multivariate Analysis 92 (2005) 186–204

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

Covariate selection for semiparametric hazard function regression models

Florentina Bunea* and Ian W. McKeague

Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA

Received 22 September 2002

Abstract

We study a flexible class of nonproportional hazard function regression models in which the influence of the covariates splits into the sum of a parametric part and a time-dependent nonparametric part. We develop a method of covariate selection for the parametric part by adjusting for the implicit fitting of the nonparametric part. Asymptotic consistency of the proposed covariate selection method is established, leading to asymptotically normal estimators of both parametric and nonparametric parts of the model *in the presence of* covariate selection. The approach is applied to a real data set and a simulation study is presented.

AMS 2000 subject classifications: 62N02

Keywords: Additive risk model; Cox model; Penalized partial likelihood; Penalized likelihood; Model selection; Survival analysis

1. Introduction

Covariate selection is a form of model selection in which the class of models under consideration is represented by subsets of covariate components to be included in the analysis. Model selection methods are well developed in parametric settings, and in recent years they have been extended to wide classes of nonparametric models [2]. For applications in survival analysis, however, in which the presence of censoring and the use of complex time-dependent hazard function regression models is

*Corresponding author.
E-mail address: flori@stat.fsu.edu (F. Bunea).

0047-259X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2003.09.006

becoming increasingly popular (see, e.g., [1]), generally applicable and fully validated procedures have not yet been developed.

In this paper we study covariate selection for conditional hazard function models of the form

$$h(t,x,z) = \psi(\beta^T x + f(t)^T z), \tag{1.1}$$

where ψ is a known (nonnegative) link function, (x,z) is a partition of the covariates into a q-vector x and a p-vector z, β is an unknown q-vector of regression parameters and $f(t)$ is an unknown p-dimensional nonrandom function of time. We develop a model selection procedure to find the best subset of x-covariates and study the asymptotic properties of the corresponding regression parameter estimates *after* model selection.

The above model provides a flexible extension of the Cox proportional hazards model $h(t,x) = \exp(\beta^T x + f(t))$, where $f(t)$ is the log-baseline hazard function. Our model is more flexible in the sense that it allows some of the covariates to have a longitudinal (or time-dependent) influence on survival. For the identity link function, the model reduces to the partly parametric additive risk model of McKeague and Sasieni [13]. Recently, Martinussen et al. [12] studied the model in the case of an exponential link function.

Typically, some covariates are known to have a longitudinal influence on survival, so those covariates are placed in z. However, only a small (but fixed) number of covariates can be treated in this way as an additional time-dependent function enters the model for each component of z. The remaining covariates are placed in x. This creates the need for a procedure to select a subset of the x-components that avoids both overfitting and underfitting. With the nonzero components of β corresponding to an unknown subset $I = I_0$ of the x-covariates, the statistical problem is to estimate I_0 and the corresponding components of β.

Numerous covariate selection procedures have been proposed for the Cox model: penalized partial likelihood—henceforth PPL [16], a backwards elimination covariate selection method [9], Bayesian model averaging [14,15], Bayesian variable selection [8], the lasso method for PPL [17], and nonconcave PPL [7]. Large sample properties of these procedures are largely unexplored, with the exceptions of Senoussi [16] and Fan and Runze [7]. All these procedures only require *parametric* model selection techniques because they exploit partial likelihood which does not involve the infinite-dimensional part of the semiparametric model (the baseline hazard function). A more sophisticated PPL procedure was developed by Letué [11] for fitting the general proportional hazards model

$$h(t,x) = \exp(g(x) + f(t)),$$

where $g(x)$ is an unknown function of the covariates x and $f(t)$ is the log-baseline hazard function. This model may be unsuitable, however, when x has high dimension because of the curse of dimensionality. None of the above procedures extends beyond the proportional hazards framework.

To study semiparametric models of form (1.1), in which a partial likelihood for β is not available and I_0 is also regarded as a parameter, we need a different approach.

We consider the following two-stage procedure. The first stage (covariate selection) is to estimate I_0 by $\hat{I}$ derived from maximizing a penalized *full likelihood*. To produce a consistent $\hat{I}$, the effect of estimating f nonparametrically needs to be controlled via a penalty that is different from the parametric model selection procedures mentioned above. In [16], for instance, the penalty term has the form $a_n|I|/n$, with $a_n \to \infty$ and $a_n/n \to 0$, where $|\cdot|$ denotes the cardinality of a set; restrictions on a_n (e.g., $a_n = \log n$) then lead to consistent estimators of I_0. This type of penalty term will not work for full likelihood because the penalty must also balance the bias caused by estimation of the infinite-dimensional part of the model; see [5] for a regression example. The second stage of our procedure is to refit the model with the x-components restricted to those in $\hat{I}$ using the estimators of β and the cumulative regression function $\int_0^t f(s)\,ds$ developed by McKeague and Sasieni [13] and Martinussen et al. [12]. The end result is consistent covariate selection along with asymptotically normal estimators of both parametric and nonparametric parts of the model.

The paper is organized as follows. In Sections 2.1 and 2.2 we introduce the proposed method. Section 2.3 contains the main result giving the consistency of $\hat{I}$ and, as an immediate consequence, the asymptotic normality of the corresponding estimator of β. In Section 3 we present a simulation study comparing the proposed approach with various competitors. An application to real data is discussed in Section 4. The proofs of intermediate results are collected in Section 5.

2. Covariate selection

In this section we present the proposed method of selecting the best subset I_0 of the x-covariates based on a penalized full likelihood procedure. The procedure leads to consistent estimates of I_0. We also establish upper bounds on the convergence rates of the corresponding estimators of β and f.

2.1. Preliminaries

The survival time T is assumed to be conditionally independent of a censoring time C given the covariates (X, Z). We observe n i.i.d. copies of the right censored survival time $T^{\text{obs}} = \min(T, C)$ and the censoring indicator $\delta = 1(T \leqslant C)$. The true conditional hazard function $h(t, X, Z)$ of T given (X, Z) is specified by (1.1) where t is restricted to a fixed time interval $[0, \tau]$. The covariates are assumed to be bounded.

We suppose that the link function ψ is positive, continuous and strictly increasing on some (sufficiently large) known bounded interval $[a, b]$ for which $\psi(a) \leqslant h(t, x, z) \leqslant \psi(b)$ for all t, x, z. This means that $h(t, x, z)$ has known uniform bounds in terms of values of the given link function. For the identity link function, $a > 0$ and b represent prespecified bounds on the hazard function; in practice, a can be chosen arbitrarily small and b arbitrarily large, so they have no effect on the estimation procedure. For the exponential link function, a and b are bounds on the

log-hazard function. The inverse of ψ is denoted $\psi^{-1}:[\psi(a),\psi(b)]\to[a,b]$. The function

$$r(t,x,z)=\psi^{-1}(h(t,x,z))=\beta^T x+f(t)^T z \tag{2.2}$$

plays a central role in our approach; in the case of the exponential link, r is simply the log-hazard function.

The following set of conditions is assumed throughout.

Conditions.

(A1) $a\leqslant r(t,X,Z)\leqslant b$.
(A2) ψ and ψ^{-1} are Lipschitz on $[a,b]$ and $[\psi(a),\psi(b)]$, respectively.
(A3) $P(C\geqslant\tau|X,Z)$ is bounded away from zero.
(A4) $\mathrm{Var}(l^T X|Z=z)>0$ for any nonzero $l\in\mathbb{R}^q$ and any z.
(A5) $\mathrm{Var}(d^T Z|X=x)>0$ for any nonzero $d\in\mathbb{R}^p$ and any x.
(A6) The components of f belong to $B^{\alpha}_{\infty}(L^2)$, for some $\frac{1}{2}<\alpha\leqslant 1$.
(A7) X and Z are uniformly bounded.

Here $B^{\alpha}_{\infty}(L^2)$ is the Besov space of order α corresponding to the L^2-space of square-integrable functions on $[0,\tau]$; see, e.g., [6] for the precise definition and properties.

Conditions (A4) and (A5) are identifiability assumptions that allow us to make separate inferences on the parametric and nonparametric parts of the model, and can be checked in practice by inspecting scatterplots of the components of X with respect to the components of Z.

2.2. Sieves and selection criterion

We now introduce suitable parametric submodels (sieves) consisting of the functions $u(t,x,z)$ that will be used to approximate the true $r(t,x,z)$. The sieves also naturally provide approximations to the true conditional hazard function.

Define the sieve

$$S_I=\bar{S}_I\cap\{u(\cdot):a\leqslant u(\cdot)\leqslant b\}$$

indexed by a given subset $I=\{i_1,\ldots,i_l\}$ of the x-covariate indices, where $\bar{S}_I$ is the finite-dimensional linear approximating space

$$\bar{S}_I=\langle x_{i_1},\ldots,x_{i_l},\phi_{n,1}(t)z_1,\ldots,\phi_{n,N_n}(t)z_1,\ldots,\phi_{n,1}(t)z_p,\ldots,\phi_{n,N_n}(t)z_p\rangle \tag{2.3}$$

with $\phi_{n,i}(t)\equiv 1_{[(i-1)/N_n,i/N_n]}(t/\tau)$, for $i=1,\ldots,N_n$, and $N_n=[n^{1/(2\alpha+1)}]$, where $[\,]$ denotes the integer part. Thus, within each S_I, the components of f are approximated by step functions based on a regular partition of $[0,\tau]$. The mesh of the partition depends on α and the sample size n. We note that although minimax adaptive estimation of these functions is possible, see, e.g, [2] for a very general approach, this would require the construction of a much richer set of approximating spaces which would result in a very involved algorithm that may become computationally intractable. Since at this stage of our estimation procedure we

only need a good initial estimate of the nonparametric part of our model, we content ourselves with the simple construction above.

To select amongst the subsets $I \subseteq \{1, \ldots, q\}$ we use the re-normalized log-likelihood as a contrast function:

$$\gamma_n(u) = -n^{-1} \sum_{i=1}^{n} \left\{ \int_0^\tau \log\psi(u(t, X_i, Z_i))\, dN_i(t) - \int_0^\tau Y_i(t)\psi(u(t, X_i, Z_i))\, dt \right\}, \tag{2.4}$$

see (6.17). The subscript i above refers to the ith individual in the sample, $N_i(t) = I(T_i^{\text{obs}} \leqslant t, \delta_i = 1)$, and $Y_i(t) = I(T_i^{\text{obs}} \geqslant t)$. We declare $\hat{r} \in S_{\hat{I}}$ a penalized maximum likelihood sieve estimator if

$$\gamma_n(\hat{r}) + \text{pen}(\hat{I}) = \inf_I \left[\inf_{u \in S_I} (\gamma_n(u) + \text{pen}(I)) \right], \tag{2.5}$$

where $\text{pen}(I)$ is a penalty term which will be defined in the next subsection.

2.3. Consistency of the selection and asymptotic normality

In this subsection we show that the method proposed above consistently estimates I_0. We indicate then how this can be used to construct asymptotically normal estimators for β.

The choice of the penalty term is crucial for this result. We begin by giving the motivation behind our choice and we defer the full proof to Theorem 2. Note that

$$P(\hat{I} \neq I_0) = P(I_0 \subsetneq \hat{I}) + P(I_0 \not\subset \hat{I}).$$

We show in Theorem 1 that our procedure leads to consistent estimators of β. Then, it is easy to show that $P(I_0 \not\subset \hat{I}) \to 0$. To see this assume that, for example, $\beta = (1, 2, 0, 3)$. Then we cannot consistently estimate it by, say, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, 0, 0)$. The study of the second inclusion is more delicate. One cannot rely on the consistency of $\hat{\beta}_{\hat{I}}$ alone to show that $P(I_0 \subsetneq \hat{I}) \to 0$, as we can always consistently estimate zeros. Thus, one needs a different argument, which will, in turn, lead to restrictions on the penalty term. Note that

$$P(I_0 \subset \hat{I}) = \sum_{I \supset I_0} P(\hat{I} = I).$$

Let $\mathbb{D}_n \equiv \mathbb{P}_n - P$, where the measure P corresponds to the hazard function $h = \psi \circ r$ and $\mathbb{P}_n$ is the empirical measure that puts mass $1/n$ at each observation. Then, as in the course of the proof of Theorem 2, by the definition of our estimator, for any $I \supset I_0$ and an appropriately defined constant B and function r_I, an upper bound for $P(\hat{I} = I)$ is given by

$$P\left(\sup_{v \in S_I} \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - ||v - r_I||_v^2 > \text{pen}(I) - \text{pen}(I_0) - BpN_n^{-2\alpha} \right). \tag{2.6}$$

Thus, a first restriction on the penalty term is

$$\text{pen}(I) - \text{pen}(I_0) > BpN_n^{-2\alpha}. \tag{2.7}$$

Note that $BpN_n^{-2\alpha}$ is a bias term introduced through the approximation of the infinite-dimensional part of the model within a space S_I of finite dimension pN_n. We emphasize that a similar derivation for a fully parametric model would *not* contain the bias term $BpN_n^{-2\alpha}$ and that a requirement of type (2.7) is not a byproduct of our method of proof, but it is intrinsic to the nature of a semiparametric model (although it can be avoided when PPL is available); see also [5] for a regression example. More generally, a penalty term satisfying (2.7) can be used in any covariate selection procedure in which the criterion γ satisfies the same bound as in (2.6). McKeague and Sasieni's [13] least-squares estimator of β applies to a model with *known* covariate structure, and avoids bias from the nonparametric part of the model. However, it would be misleading to conclude that covariate selection can avoid bias from the nonparametric part in this model: beyond the PPL framework it is necessary to use the full likelihood and take the bias into account.

For our choice of N_n and with $|I|$ denoting the cardinality of I, the following penalty term

$$\text{pen}(I) = C\frac{(|I|+1)pN_n}{n}\log n, \tag{2.8}$$

satisfies (2.7), for n large enough. We discuss the choice of C for large and small samples in Section 5.1. Notice that a penalty term in which the dimensions of the two parts of the approximating space are added rather than multiplied does not satisfy (2.8); thus, a direct extension of the penalty terms developed for selection methods in which the parameter of interest belongs to one of the candidate spaces fails in this context. For this choice of penalty, we prove in Theorem 2 that, asymptotically, we cannot underestimate or overestimate I_0. As we discussed above, to show that we cannot underestimate I_0, we shall first show that the selected $\hat{\beta}$ is consistent. We prove this in Corollary 1, which is an immediate consequence of the much stronger result of Theorem 1, in which we give finite sample upper bounds on the risk of our estimators.

Let r_I denote the orthogonal projection of r onto $\bar{S}_I$ in the L_v^2 space corresponding to the measure $v = \text{Leb} \times \mu_{X,Z}$ on $[0,\tau] \times \mathbb{R}^q \times \mathbb{R}^p$ and $\mu_{X,Z}$ is the distribution of (X,Z). Notice that, by Lemma 3, $r_I \in S_I$. The L_v^2 and L_{Leb}^2-norms, respectively, are denoted $||\cdot||_v$ and $||\cdot||_2$. The Euclidean norm is denoted $|\cdot|_2$.

Theorem 1. *Under Conditions* (A1)–(A7), *for the estimators relative to the collection of approximating spaces* (2.3) *and for the penalty term* (2.8) *there exist positive constants* C_1 *and* C_2 *such that*

$$E_P||r - \hat{r}||_v^2 \leqslant C_1 \inf_I [||r - r_I||_v^2 + \text{pen}(I) + C_2/n]. \tag{2.9}$$

The following corollary provides rates of convergence of the estimators $\hat{\beta}$ and $\hat{f}$ corresponding to $\hat{r}$. Let $\mathscr{D}_\alpha$ denote the family of functions $f=(f_1,\ldots,f_p)^T$ with each component f_j belonging to a fixed bounded subset of the Besov space in Condition (A6).

Corollary 1. *Under the conditions of Theorem* 1, *we obtain*

1. $||\hat{f}_j - f_j||_2^2 = O_P(\frac{\log n}{n^{2\alpha/(2\alpha+1)}})$, *uniformly over* $f \in \mathscr{D}_\alpha$.
2. $|\hat{\beta} - \beta|_2^2 = O_P(\frac{\log n}{n^{2\alpha/(2\alpha+1)}})$, *uniformly over* β *in any compact* $\mathscr{K} \subset \mathbb{R}^q$.

In the above result, the rate for $\hat{f}$ is the minimax optimal nonparametric rate, up to a $\log n$ factor, but the rate for $\hat{\beta}$ is not the optimal $\sqrt{n}$-rate. For this we need the following consistency result for $\hat{I}$.

Theorem 2. *Under the conditions of Theorem* 1, *we have* $P(\hat{I} \neq I_0) \to 0$ *as* $n \to \infty$.

Proof. Note that

$$P(\hat{I} \neq I_0) = P(I_0 \subsetneq \hat{I}) + P(I_0 \not\subset \hat{I}). \tag{2.10}$$

We show that each term in the right-hand side of (2.10) converges to zero.

1. $P(I_0 \subsetneq \hat{I}) \to 0$ as $n \to \infty$.

Notice that if $I_0 = \{1,\ldots,q\}$, $P(I_0 \subsetneq \hat{I}) = 0$, so it is enough to consider $I_0 \subsetneq \{1,\ldots,q\}$.

We can write

$$P(I_0 \subset \hat{I}) = \sum_{I \supset I_0} P(\hat{I} = I), \tag{2.11}$$

where $I \subset \{1,\ldots,q\}$. Define

$$f_n(I) \equiv \inf_{u \in S_I} (\gamma_n(u) + \mathrm{pen}(I)). \tag{2.12}$$

By the definition of the estimator we have

$$\begin{aligned} P(\hat{I} = I) &= P(f_n(I) - f_n(I') < 0, \text{ for all } I' \neq I) \\ &\leqslant P(f_n(I) - f_n(I_0) < 0). \end{aligned} \tag{2.13}$$

With notation (1.12), by adding and subtracting $\gamma_n(r_I) + \mathrm{pen}(I_0)$, we have

$$\begin{aligned} f_n(I) - f_n(I_0) = &- \sup_{v \in S_{m(I)}} [\gamma_n(r_I) - \gamma_n(v) - \mathrm{pen}(I) + \mathrm{pen}(I_0)] \\ &- \inf_{u \in S_{I_0}} [\gamma_n(u) - \gamma_n(r_I) + \mathrm{pen}(I_0) - \mathrm{pen}(I_0)]. \end{aligned}$$

By (2.11), we restrict attention only to $I \supset I_0$. By Lemma 3, $r_I \in S_I$ for any $I \supset I_0$. Then

$$\begin{aligned}
&P(f_n(I) - f_n(I_0) < 0) \\
&\quad < P\left(\sup_{v \in S_I} [\gamma_n(r_I) - \gamma_n(v) - \mathrm{pen}(I) + \mathrm{pen}(I_0)] > 0 \right) \\
&\qquad + P\left(\inf_{u \in S_{I_0}} [\gamma_n(u) - \gamma_n(r_I)] > 0 \right) \\
&\quad < P\left(\sup_{v \in S_I} [\gamma_n(r_I) - \gamma_n(v) - \mathrm{pen}(I) + \mathrm{pen}(I_0)] > 0 \right) \\
&\qquad + P(\gamma_n(r_I) - \gamma_n(r_I) > 0) \\
&\quad = P\left(\sup_{v \in S_I} [\gamma_n(r_I) - \gamma_n(v) - \mathrm{pen}\ (I) + \mathrm{pen}\ (I_0)] > 0 \right). \qquad (2.14)
\end{aligned}$$

Notice now that, for c_1 and c_2 given by Lemma 1 in Section 5, we obtain

$$\begin{aligned}
\gamma_n(r_I) - \gamma_n(v) &= \mathbb{D}_n[\gamma(r_I) - \gamma(v)] + E_P(\gamma(r_I) - \gamma(r)) - E_P(\gamma(v) - \gamma(r)) \\
&\leqslant \mathbb{D}_n[\gamma(r_I) - \gamma(v)] + c_2 ||r_I - r||_v^2 - c_1 ||v - r||_v^2 \\
&\leqslant \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - c_1 ||v - r_I||_v^2/2 + (c_1 + c_2)||r_I - r||_v^2 \\
&\leqslant \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - c_1 ||v - r_I||_v^2/2 + B_1(c_1 + c_2) n^{-2\alpha/(2\alpha+1)},
\end{aligned}$$

since $||v - r_I||_v^2 \leqslant 2(||v - r||_v^2 + ||r_I - r||_v^2)$ and for B_1 given in the proof of Corollary 1. Then, with $\mathrm{pen}(I) = C(|I| + 1)n^{-2\alpha/(2\alpha+1)} \log n$, for n large enough and a dominating constant L^*, we obtain

$$\begin{aligned}
&\gamma_n(r_I) - \gamma_n(v) - \mathrm{pen}(I) + \mathrm{pen}(I_0) \\
&\quad \leqslant \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - c_1 ||v - r_I||_v^2/2 + B_1(c_1 + c_2) n^{-2\alpha/(2\alpha+1)} - C n^{-2\alpha/(2\alpha+1)} \log n \\
&\quad \leqslant \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - c_1 ||v - r_I||_v^2/2 - L^* n^{-2\alpha/(2\alpha+1)} \log n.
\end{aligned}$$

Notice now that $2n^{1/(2\alpha+1)} \geqslant (|I| + p n^{1/(2\alpha+1)})/p(q+1)$. Let

$$\sigma_n \equiv (|I| + p n^{1/(2\alpha+1)}) \log n / 2np(q+1).$$

Let $A \equiv L^*/2p(q+1)$ and $A^* \equiv \min(c_1/2, A)$. Then, noting that $|I| + p[n^{1/(2\alpha+1)}]$ is the dimension of S_I, we apply Theorem 5 of Brigé and Massart [3]. For positive constants C_3 and C_4 given by this theorem, we obtain, for

any $I \supset I_0$, that

$$P\left(\sup_{v \in S_I} [\gamma_n(r_I) - \gamma_n(v) - \mathrm{pen}(I) + \mathrm{pen}(I_0)] > 0\right)$$

$$\leqslant P\left(\sup_{v \in S_I} \mathbb{D}_n[\gamma(r_I) - \gamma(v)] > A^*[||r_n - v||_v^2 + \sigma_n]\right)$$

$$\leqslant P\left(\sup_{v \in S_I} \frac{\mathbb{D}_n[\gamma(r_I) - \gamma(v)]}{||r_n - v||_v^2 \vee \sigma_n} > 1/A^*\right) \leqslant C_3 \exp(-C_4 n\sigma_n) \to 0. \tag{2.15}$$

Note that the hypotheses of Theorem 5 of [3] are verified in the course of the proof of our Theorem 1. Then, from (2.11)–(2.15), we have $P(I_0 \subset \hat{I}) \to 0$, which completes the proof of this step.

2. $P(I_0 \not\subset \hat{I}) \to 0$.

$$P(I_0 \not\subset \hat{I}) = P(j \notin \hat{I} \text{ for all } j \in I_0) \leqslant P(j_0 \notin \hat{I} \text{ for some } j_0 \in I_0)$$

$$\leqslant P(j_0 \in I_0 - \hat{I}) \leqslant P(\beta_{j_0} \neq 0, \hat{\beta}_{j_0} = 0)$$

$$\leqslant P(|\hat{\beta}_{j_0} - \beta_{j_0}| = |\beta_{j_0}| > 0) \to 0,$$

by the component-wise consistency of $\hat{\beta}$. This completes the proof of this theorem. □

Once the consistency of $\hat{I}$ has been established, we can refit model (1.1) with the X covariates corresponding to the index set $\hat{I}$. For this stage of our procedure, any method that leads to asymptotically normal estimators of β in (1.1), for *known* I_0, can be employed. Let $\tilde{\beta}$ be the generic notation for an estimator obtained through such method. Let $q_0 \equiv |I_0|$. Let $\Sigma^0 \equiv (\sigma^0_{i,j})_{(i,j) \in I_0 \times I_0}$ be a $q_0 \times q_0$ positive definite matrix. Let Σ be a $q \times q$ positive definite matrix, with $\sigma_{i,j} = \sigma^0_{i,j}$, for $(i,j) \in I_0 \times I_0$, and zero otherwise. Also, we denote by β_0 the nonzero components of β, so that $\beta_0 \in \mathbb{R}^{q_0}$, and we denote by $\tilde{\beta}_0 \in \mathbb{R}^{q_0}$ its estimator. In order to emphasize the possible change in dimension, we shall denote by $\tilde{\beta}_{\hat{I}} \in \mathbb{R}^q$ the estimator in the model with $\hat{I}$ covariates, to which we added zero's to the necessary positions.

Theorem 3. *Let $\tilde{\beta}$ be any estimator such that*

$$\sqrt{n}(\tilde{\beta}_0 - \beta_0) \to_d N_{q_0}(0, \Sigma_0).$$

Under Conditions (A1)–(A7) *we then have*

$$\sqrt{n}(\tilde{\beta}_{\hat{I}} - \beta) \to_d N_q(0, \Sigma),$$

where the limiting distribution has all its mass concentrated on the space generated by the true I_0 covariates.

The proof of this theorem is identical to the proof of Theorem 5.2 of Bunea [2]. It relies on the fact that

$$P\big(\sqrt{n}c^T(\tilde{\beta}_{\hat{I}}-\beta)\leqslant b\big)=P\big(\sqrt{n}c^T(\tilde{\beta}_{\hat{I}}-\beta)\leqslant b,\hat{I}=I_0\big)+P\big(\sqrt{n}c^T(\tilde{\beta}_{\hat{I}}-\beta)\leqslant b,\hat{I}\neq I_0\big),$$

for any $c\in\mathbb{R}^q$, $b\in\mathbb{R}$. Then, one uses Theorem 2 and the asymptotic normality for fixed I_0 to establish the result.

This result validates the following selection–estimation strategy.

Step 1: estimate I_0 by $\hat{I}$.

Step 2: use an estimation method that yields an asymptotically normal estimator of β_0 under *known* I_0, replacing I_0 by $\hat{I}$.

Theorem 3 then guarantees that the resulting estimator of β is $\sqrt{n}$-consistent and asymptotically normal. The iterative estimation method of Martinussen et al. [12, Section 3] can be adapted for Step 2, and also to provide a consistent estimate of the asymptotic covariance matrix. Referring to their model formulation, we now briefly indicate (in *their* notation) the changes that are needed to extend the iteratively defined estimators to our more general model (1.1). Their exponential link function is replaced by ψ, here assumed to be differentiable, and the derivative of $\log\psi$ (identically 1 for the exponential link) is denoted κ. Their matrices $X(t)$ and $Z(t)$ are replaced, respectively, by $(\tilde{x}_1(t),\ldots,\tilde{x}_n(t))$ and $(\tilde{z}_1(t),\ldots,\tilde{z}_n(t))$, where

$$\tilde{x}_i(t)=\kappa(x_i^T\beta(t)+z_i^T\gamma)x_i,\quad \tilde{z}_i(t)=\kappa(x_i^T\beta(t)+z_i^T\gamma)z_i.$$

Note that their γ is β_0 in our notation, and their $B(t)=\int_0^t f(s)\,ds$ in our notation. This leads to asymptotically normal (and efficient) estimators of the parametric and nonparametric parts of model (1.1) in the presence of consistent covariate selection.

3. Simulation study

This section reports some simulation results designed to compare the proposed approach with various competitors.

We compare our penalized full likelihood (proposed-PFL) procedure with the penalized partial likelihood (PPL) procedure having penalty term $\mathrm{pen}(I)=C\log n|I|/n$, as used by Senoussi [16]. Note that, in the special case of the Cox model, this is a comparison between two asymptotically consistent methods. We also compare with the performance of an alternative PFL procedure (naive-PFL) having penalty term $\mathrm{pen}(I)=C\log n(|I|+pN_n)/n$, which may be regarded as a naive adjustment for the bias caused by estimating the nonparametric part of the model. To give a fair comparison between the three procedures we restrict our simulation study to the case of a Cox model (exponential link function and $p=1$).

The data were simulated using the conditional hazard function (1.1) with exponential link, $q = 7$, $\beta = (1, 1, 0, 0, 0, 1, 0)^T$, $p = 1$ and $f(t) \equiv 0$, which is a Cox model with constant baseline hazard function, so both PPL and proposed-PFL are consistent. The covariates X were i.i.d. uniform $(0, 1)$, and $Z \equiv 1$. The censoring time was taken as exponential with rate 0.5, and the end of follow-up $\tau = 0.3$. Note that the true x-covariate indices are $I_0 = \{1, 2, 6\}$.

We began by simulating 50 datasets each with 50 observations and estimated β via each procedure. We chose $\alpha = 1$ in the case of proposed-PFL. To calibrate the tuning constant C in each case, we varied C over a fine grid and examined boxplots of the estimates of β; the best results were obtained with $C = 0.7$ for proposed-PFL and $C = 0.3$ for the other two procedures, see the left panels of Fig. 1. All three procedures perform well and there is no clear winner at this sample size. We note that a choice of $C \geqslant 1$ leads to underfitting, with more relevant covariates being left out as we increased C, whereas $C \leqslant 0.2$ leads to overfitting. In agreement with the simulation results in [4], we conjecture that in practice a constant $C < 1$ needs to be used, large values of C leading to major underfitting, especially for small and moderate (less than 500) sample sizes. Next we simulated 50 datasets each with 200 observations and applied the procedures using the tuning constant C calibrated for $n = 50$; see the right panels of Fig. 1. It is clear that the proposed procedure outperforms both PPL and naive-PFL, and correctly identifies the zero coefficients of β in almost every case. The results strongly suggest that our approach achieves consistency of $\hat{I}$ at a faster rate than both the PPL and naive-PFL methods.

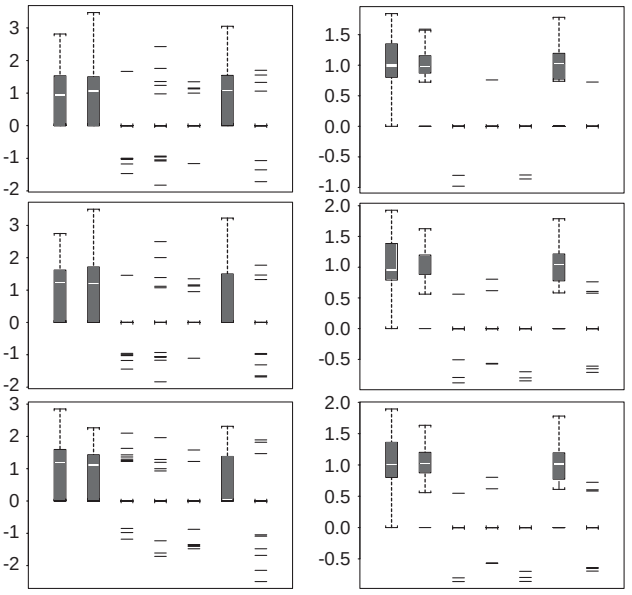

Fig. 1. Componentwise boxplots of estimates of β. Top row: proposed-PFL ($n = 50, 200$). Middle row: PPL ($n = 50, 200$). Bottom row: naive-PFL ($n = 50, 200$).

4. Example

The data come from a Mayo Clinic trial in primary biliary cirrhosis of the liver, see [9]. Times between registration and death (possibly right censored) are available for 312 patients; we only consider the 276 patients for whom complete covariate information is available at registration. Nine of the 17 covariates clearly have no effect and are excluded. We restrict attention to the following eight:

age	age in yr
edema	presence of edema (0 = no, 0.5 resolved, 1 = unresolved with therapy)
bili	serum bilirubin, in mg/dl
albu	albumin, in gm/dl
copp	urine copper, in μg/day
SGOT	SGOT, in U/ml
thromb	prothrombine time, in s
hist	histologic stage of disease, graded 1, 2, 3, or 4

Of these covariates, bili, albu and thromb were log-transformed. We used an exponential link, $\tau = 3500$ d, $\alpha = 1$, $p = 1$, $Z \equiv 1$, and the same penalty constants C as in the simulation study above. The results are displayed in Table 1.

We find the best subset of covariates to be $\hat{I} = \{\text{bili}, \text{albu}, \text{copp}, \text{thromb}, \text{hist}\}$ using the proposed-PFL method. In contrast, the PPL method gives $\{\text{age}, \text{bili}, \text{albu}, \text{copp}, \text{hist}\}$, and this is in essential agreement with the lasso solution of Tibshirani [17]. On the other hand, Fleming and Harrington [9], using a backwards elimination method, concluded that the best selection was $\{\text{age}, \text{edema}, \text{bili}, \text{albu}, \text{thromb}\}$, as did Raftery et al. [14] using Bayesian model averaging.

Our approach yields a different result to these previous analyses. In particular, the variable thromb has a high Z-score under our procedure, in marked disagreement with PPL which does not include that covariate in the selected model.

Table 1
Comparison of estimates of β for the Mayo Clinic data

Covariate	PPL			Naive-PFL			Proposed-PFL		
	Coeff	SE	Z-score	Coeff	SE	Z-score	Coeff	SE	Z-score
age	0.03	0.009	2.92	0.02	0.009	2.16	0.	—	—
edema	0.98	0.39	3.08	1.12	0.33	3.39	1.05	0.32	3.21
bili	0.78	0.11	6.67	0.76	0.12	6.32	0.75	0.12	6.23
albu	−2.25	0.83	−2.71	−3.58	0.74	−4.81	−3.77	0.73	−5.12
copp	0.002	0.001	2.36	0.002	0.001	2.46	0.003	0.001	2.65
SGOT	0.	—	—	0.	—	—	0.	—	—
thromb	0.	—	—	−3.08	0.62	−4.90	−2.63	0.59	−4.42
hist	0.33	0.14	2.28	0.35	0.15	2.28	0.41	0.15	2.67

The naive-PFL method only excludes one of the covariates (SGOT), and on the basis of the simulation study we may explain this as overfitting due to the incorrect penalty term.

5. Discussion

This paper presents a new method of estimation in semiparametric hazard function regression models. Although much work has been done on estimation of models of *known* parametric dimension, very little exists on estimating this dimension. Furthermore, in the semiparametric setting, there is no established methodology that is tailored to this situation. In this paper we have bridged this gap and established the theoretical properties of our suggested estimators, in a way that allows post-model selection inference. We emphasize that the form of the penalty term is a crucial ingredient in achieving optimal estimators, and that this form is not merely an extension of the penalizations used in parametric models or the Cox model. The validation of our procedure is given in Theorems 2 and 3. We note that the second step of our procedure is flexible, and it does not create computational difficulties, as any established efficient algorithm for estimation in models of *known* dimension can be used at this stage. The possible drawback of our procedure is that at its first stage, we must search through a large model space, if q is very large. However, for $q \leqslant 15$, computation is feasible, and that is typically the case for the medical applications of this model. Further investigation of computational issues is beyond the scope of this paper, and it is the subject of future research.

6. Proofs

We first give some counting process notation used in the proofs. Let $N(t) = I(T^{\text{obs}} \leqslant t, \delta = 1)$ be the single-jump counting process that registers whether an uncensored failure has occurred by time t, and $Y(t) = I(T^{\text{obs}} \geqslant t)$ the corresponding "at risk" indicator. Define the filtration $\mathscr{F}_t = \mathscr{F}_0 \vee \sigma\{N(s): s \leqslant t\}$, where $\mathscr{F}_0 = \sigma(X, Z)$. Under the true probability measure P on $\mathscr{F} \equiv \mathscr{F}_\tau$, the counting process $N(t)$ has intensity process $\lambda(t) = Y(t)h(t, X, Z)$, which means that

$$M(t) = N(t) - \int_0^t \lambda(s)\, ds, \quad t \in [0, \tau] \tag{6.16}$$

is an $\mathscr{F}_t$-martingale under P.

If the conditional hazard function changes from h to h' and the distribution of the covariates is unchanged, then we write the new probability measure on $\mathscr{F}$ as P'. The intensity of N under P' is then $\lambda'(t) = Y(t)h'(t, X, Z)$. If $h(t, X, Z)$ and $h'(t, X, Z)$ are bounded and bounded away from zero over $t \in [0, \tau]$ a.s., then the restrictions P'_t and P_t of P' and P to $\mathscr{F}_t$ are mutually absolutely continuous and the

log-likelihood ratio is

$$\log \frac{dP'_t}{dP_t} = \int_0^t \log\left(\frac{\lambda'(s)}{\lambda(s)}\right) dN(s) - \int_0^t (\lambda'(s) - \lambda(s))\, ds, \tag{6.17}$$

where $\log 0/0 = 0$, see Andersen et al. (1993, p. 98).

The Hellinger distance between two probability measures P and P' is defined by

$$\rho^2(P, P') = \tfrac{1}{2} E_Q\left(\sqrt{V} - \sqrt{V'}\right)^2,$$

where $Q = (P + P')/2$, $V = dP/dQ$, and $V' = dP'/dQ$. Note that $\rho^2(P, P')$ does not depend on the choice of the dominating measure Q. The Kullback–Liebler information number between P and P' is $K(P, P') = \int \log(dP/dP')\, dP$ when P is absolutely continuous with respect to P', otherwise $K(P, P') = \infty$.

6.1. Proof of Theorem 1

It suffices to verify conditions C, p. 377, and M (6.4)–(6.6), pp. 371–372, of Theorem 8, p. 378, of [2] Barron et al. [2] in our context.

Lemma 1 provides the "closing argument" of Barron et al. (condition C, p. 377), and gives an equivalence (up to constants) between the Kullback–Liebler information number

$$K(P, P') = E(\gamma_n(u) - \gamma_n(r))$$

and the L_v^2-norm between u and r (c.f., [3, Section 4.2]). This equivalence is established using a result of Jacod and Shiryaev [10] which allows us to express the Hellinger distance between two counting processes in terms of their intensities.

In Lemma 2 we check condition M (6.4). This amounts to checking a Lipschitz type condition on the process $u \mapsto \gamma(u)$:

$$\gamma(u) = \gamma(u, T^{\text{obs}}, \delta, X, Z) = \int_0^\tau Y(t)\psi(u(t, X, Z))\, dt - \int_0^\tau \log \psi(u(t, X, Z))\, dN(t). \tag{6.18}$$

Condition M (6.5) holds by Lemma 9, p. 400 of Barron et al. [2, p. 372]. Assumption M (6.6) follows by the definition of γ in (6.18), our Conditions A and the Lipschitz property of log on $[\psi(a), \psi(b)]$. Then, for some constant $a_1 > 0$, we have

$$||\Delta(\cdot, u, v)||_\infty \leqslant a_1 ||v - u||_\infty.$$

Lastly, note that Theorem 8 of Barron at al. [2] applies for any penalty term greater or equal than $\tilde{C}(|I| + pN_n)/n$, where $|I| + pN_n$ is the dimension of the approximating space and $\tilde{C}$ is a positive constant given by their Theorem. Notice that, since $|I| + 1, pN_n \geqslant 1$, then $2(|I| + 1)pN_n \geqslant (|I| + pN_n)$ and so $2(|I| + 1)pN_n \log n/n > \tilde{C}(|I| + pN_n)/n$, for n large enough. Thus, in the definition of our penalty term, one can take $C = 2$. However, other choices are possible, and

for small sample sizes it is easy to calibrate the value of C via simulation, as we did in Section 3.

Lemma 1. *Suppose $u(t,x,z)$ satisfies Condition* (A1) *in place of r. Let P and P' correspond to the conditional hazard functions $h=\psi\circ r$ and $h'=\psi\circ u$, respectively. Then there exist constants $0<c_1<c_2$ such that*

$$c_1||r-u||_v^2 \leqslant K(P,P') \leqslant c_2||r-u||_v^2.$$

Proof. First note that $P_0=P_0'$, so $\rho^2(P_0,P_0')=0$. Proposition 1.27 and Theorem 4.2 of Jacod and Shiryaev [10, p.197, 237)], applied with $Q=P$, then give

$$\rho^2(P,P')=\frac{1}{2}E_P\int_0^\tau \sqrt{V'_{s-}}\left(\sqrt{\lambda(s)}-\sqrt{\lambda'(s)}\right)^2 ds,$$

where $V'_t=dP'_t/dP_t$ is given by (6.17). Using (6.17), the bounds on r and u and the fact that N has at most a single jump, it can be easily seen that V'_t is bounded and bounded away from zero by constants that only depend on $a,\ b,\ \varepsilon$ and τ. Thus, in the sense of the conclusion of the lemma,

$$\rho^2(P,P')\simeq E_P\int_0^\tau\left(\sqrt{\lambda(s)}-\sqrt{\lambda'(s)}\right)^2 ds. \tag{6.19}$$

By Birgé and Massart [3, (7.5), (7.6)] we have

$$2\rho^2(P,P')\leqslant K(P,P')\leqslant \rho^2(P,P')(4+2\log||V||_\infty) \tag{6.20}$$

where $||V||_\infty$ is the supremum norm of $V=dP/dP'$. Combining (6.19) and (6.20) we find that

$$K(P,P')\simeq E_P\int_0^\tau\left(\sqrt{\lambda(s)}-\sqrt{\lambda'(s)}\right)^2 ds. \tag{6.21}$$

Using the fact that C is conditionally independent of T given the covariates (X,Z), as well as the upper bound $\psi(b)$ on h and the lower bound on $P(C\geqslant t|X,Z)$ in Condition (A3), we have

$$E_P(Y(t)|X,Z)=P(C\geqslant t|X,Z)E_P(T\geqslant t|X,Z)\geqslant \varepsilon\exp(-\tau\psi(b))$$

almost surely, for some $\varepsilon>0$, so conditioning on (X,Z) we find that

$$\begin{aligned}
E_P\int_0^\tau\left(\sqrt{\lambda(s)}-\sqrt{\lambda'(s)}\right)^2 ds &= E_P\int_0^\tau\left(\sqrt{h(s,X,Z)}-\sqrt{\psi(u(s,X,Z))}\right)^2 Y(s)\,ds\\
&\simeq E_P\int_0^\tau\left(\sqrt{h(t,X,Z)}-\sqrt{\psi(u(t,X,Z))}\right)^2 dt\\
&\simeq E_P\int_0^\tau(\psi(r(t,X,Z))-\psi(u(t,X,Z)))^2\,dt\\
&\simeq ||r-u||_v^2.
\end{aligned}$$

The penultimate line above follows from the bounds on r and u, and the Lipschitz property of the square-root function on $[\psi(a),\psi(b)]$, where here $\psi(a)>0$. The last

line above follows from the Lipschitz assumptions on ψ and ψ^{-1}. This combined with (6.21) completes the proof. □

Lemma 2. *Suppose $u(t,x,z)$ and $v(t,x,z)$ satisfy Condition* (A1) *in place of r. Then there exists a constant $c_3>0$ only depending on a, b and τ such that*

$$E_P(\gamma(u)-\gamma(v))^2 \leqslant c_3||u-v||_v^2.$$

Proof. First note that, in terms of the martingale (6.16), we have

$$\begin{aligned}\gamma(u)-\gamma(v) &= \int_0^\tau Y(t)[\psi(u(t,X,Z))-\psi(v(t,X,Z))]\,dt \\ &\quad - \int_0^\tau \log\frac{\psi(u(t,X,Z))}{\psi(v(t,X,Z))}\,dM(t) \\ &\quad - \int_0^\tau \log\frac{\psi(u(t,X,Z))}{\psi(v(t,X,Z))}\,Y(t)h(t,X,Z)\,dt. \end{aligned} \tag{6.22}$$

Consider the first term on the r.h.s. above. By the Cauchy–Schwarz inequality applied to the integral over t, we have

$$E_P\left(\int_0^\tau Y(t)[\psi(u(t,X,Z))-\psi(v(t,X,Z))]\,dt\right)^2 \leqslant \tau||\psi\circ u-\psi\circ v||_v^2.$$

Using standard results on martingale integrals (see, e.g., [1, p. 78]), the second term on the r.h.s. in (6.22) has second moment

$$\begin{aligned}E_P\left(\int_0^\tau \log\frac{\psi(u(t,X,Z))}{\psi(v(t,X,Z))}\,dM(t)\right)^2 &= E_P\int_0^\tau\left(\log\frac{\psi(u(t,X,Z))}{\psi(v(t,X,Z))}\right)^2 Y(t)h(t,X,Z)\,dt \\ &\leqslant c_4||\psi\circ u-\psi\circ v||_v^2\end{aligned}$$

for some constant $c_4>0$, where the last inequality uses the bounds on u, v, h, and the Lipschitz property of log on $[\psi(a),\psi(b)]$. The second moment of the last term in (6.22) can be handled in a similar way to the first. The result now follows by applying the inequality $(x+y+z)^2 \leqslant 3(x^2+y^2+z^2)$ to the square of the r.h.s. of (6.22), using the second moment bounds already established and the Lipschitz assumption on ψ. □

6.2. *Proof of Corollary 1*

We begin by giving some properties of the orthogonal projection operators that we consider.

Lemma 3. *Let $r_I \equiv \pi(r)$ denote the orthogonal L_v^2-projection of r onto $\bar{S}_I$. Let $f_{j,n} = \pi_n(f_j)$, where π_n denotes the L^2_{Leb}-projection onto $\langle \phi_{n,i}\colon i=1,\ldots,N_n\rangle$, $j=1,\ldots,p$.*

(i) *$r_I(t,x,z) = \beta^T x + \sum_{j=1}^p z_j f_{j,n}(t)$, for any $I \supseteq I_0$.*
(ii) *Under* (A1), *$r_I \in S_I$, for any I.*

Proof. (*i*) By the linearity of π and since $\beta^T x \in S_I$, for any $I \supseteq I_0$, $r_I(t,x,z) = \beta^T x + \sum_{j=1}^p \pi(z_j f_j(t)) = \beta^T x + \sum_{j=1}^p z_j(\pi_n f_j)(t)$. Hence, we need to show that $\pi(z_j f_j(t)) = z_j(\pi_n f_j)(t)$.

Recall from the uniqueness of the (Riesz) orthogonal decomposition that for any $g \in L_v^2$, $\pi(g)$ is the unique element in $\bar{S}_I$ such that $g - \pi(g)$ is orthogonal to $\bar{S}_I$ (note that S_I is finite dimensional and thus closed).

By applying the above property to $g(t,x,z) = z_j f_j(t)$ we see that it suffices to show that the L_v^2-inner product between $z_j f_j(t) - z_j(\pi_n f_j)(t)$ and each generating function in $\bar{S}_I$ is zero. But the generator of the form $z_k \phi_{n,l}(t)$ has L_v^2-inner product with $z_j f_j(t) - z_j(\pi_n f_j)(t)$ given by (where we separate the variables using Fubini's theorem)

$$E_P(Z_j Z_k) \times \int_0^\tau (f_j(t) - (\pi_n f_j)(t))\phi_{n,l}(t)\, dt.$$

Notice that the second factor above is the L^2_{Leb}-inner product, so it vanishes by the orthogonal decomposition for the projection π_n. The same argument works for the generators of the form x_k.

(ii) Notice that by the argument used in (i) and by regarding $r = \beta^T x + f^T z$ as a function of t with x and z fixed, we have

$$r_I = \pi(r) = \beta^T x + \pi_n(f\,)^T z = \pi_n(\beta^T x + f^T z) = \pi_n(r).$$

Then, since the projection operator π_n is order preserving and, by (A1), $a \leqslant r \leqslant b$, we have $a \leqslant r_I \leqslant b$, which completes the proof of this lemma. □

Proof of Corollary 1. First note that for $f \in \mathscr{D}_\alpha$ the bound on the approximation error given by DeVore and Lorentz [6, Theorem 2.4, p. 358] is $||f_j - f_{j,n}||_2^2 \leqslant B(\alpha) n^{-2\alpha/(2\alpha+1)}$, for a constant $B(\alpha) > 0$ given by their theorem. By Lemma 3, for any $I \supseteq I_0$, the L_v^2-projection of r onto $\bar{S}_I$ is $r_I(t,x,z) = \beta^T x + \sum_{j=1}^p z_j f_{j,n}(t)$ and $r_I \in S_I$. Then

$$\begin{aligned} ||r - r_I||_v^2 &= \left|\left|\sum_{j=1}^p z_j(f_j - f_{j,n})\right|\right|_v^2 \leqslant pM^2 \sum_{j=1}^p ||f_j - f_{j,n}||_2^2 \\ &\leqslant p^2 M^2 B(\alpha) n^{-2\alpha/(2\alpha+1)} \equiv B_1 n^{-2\alpha/(2\alpha+1)}, \end{aligned}$$

where M is a uniform bound on the absolute value of the components of Z.

Hence, by Theorem 1, for a dominating constant $C^* > 0$ (depending on $\alpha, B, a, b, \varepsilon, \tau, M, p$), we have

$$\begin{aligned} E_P||r - \hat{r}||_v^2 &\leqslant C_1 \inf_{I \supseteq I_0} (||r - r_I||_v^2 + \text{pen}(I) + C_2/n) \\ &\leqslant C_1 \left(\frac{B_1}{n^{2\alpha/(2\alpha+1)}} + \frac{C(|I_0|+1)p}{n^{2\alpha/(2\alpha+1)}} \log n \right) \leqslant C^* \frac{\log n}{n^{2\alpha/(2\alpha+1)}}. \end{aligned} \tag{6.23}$$

Let now $(\tilde{X}, \tilde{Z}) \sim \mu_{X,Z}$, with $(\tilde{X}, \tilde{Z})$ independent of $(X_1, Z_1), \ldots, (X_n, Z_n)$ and notice that $\tilde{X} - E(\tilde{X}|\tilde{Z}) \perp_{\mu_{X,Z}} \tilde{Z}$. Then, writing E_v for integration w.r.t. v,

by Pythagoras

$$\begin{aligned}||r-\hat{r}||_v^2 &= E_v(r-\hat{r})^2(t,\tilde{X},\tilde{Z})\\ &= E_v[(f-\hat{f})^T\tilde{Z}+(\hat{\beta}-\beta)^T E_v(\tilde{X}|\tilde{Z})+(\hat{\beta}-\beta)^T(\tilde{X}-E_v(\tilde{X}|\tilde{Z}))]^2\\ &= E_v[(f-\hat{f})^T\tilde{Z}+(\hat{\beta}-\beta)^T E(\tilde{X}|\tilde{Z})]^2\\ &\quad + E_v[(\hat{\beta}-\beta)^T(\tilde{X}-E(\tilde{X}|\tilde{Z}))]^2.\end{aligned}$$

Since $||r-\hat{r}||_v^2 = O_P(\frac{\log n}{n^{2\alpha/(2\alpha+1)}})$ by (6.23), we have

$$E_v[(\hat{\beta}-\beta)^T(\tilde{X}-E(\tilde{X}|\tilde{Z}))]^2 = O_P\left(\frac{\log n}{n^{2\alpha/(2\alpha+1)}}\right). \tag{6.24}$$

Define $\Sigma=(\sigma_{ij})_{q\times q}$, with

$$\begin{aligned}\sigma_{ij} &= \mathrm{Cov}(X_i-\theta_i(Z),X_j-\theta_j(Z))\\ &= \mathrm{Cov}(X_i,X_j)-\mathrm{Cov}(\theta_i(Z),\theta_j(Z)),\end{aligned} \tag{6.25}$$

for $\theta_i(z)\equiv E_v(X_i|Z=z)$. Since $(\tilde{X},\tilde{Z})$ is independent of $\hat{\beta}$ by construction, we also have

$$\begin{aligned}E_v[(\hat{\beta}-\beta)^T(\tilde{X}-E(\tilde{X}|\tilde{Z}))]^2 &= (\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)\\ &\geqslant \lambda_{\min}|\hat{\beta}-\beta|_2^2,\end{aligned} \tag{6.26}$$

where we denoted by $\lambda_{\min}$ the smallest eigenvalue of Σ. Since, by Condition (A4), for any nonzero $l\in\mathbb{R}^q$

$$l'\Sigma l = \mathrm{Var}(l'(X-E(X|Z)) = \int \mathrm{Var}(l'X|z)\,d\mu_Z(z) > 0, \tag{6.27}$$

Σ is positive definite and so $\lambda_{\min}>0$. Thus, from (6.24) and (6.26), we have that for any $\beta\in\mathscr{K}$

$$|\hat{\beta}-\beta|_2^2 = O_P\left(\frac{\log n}{n^{2\alpha/(2\alpha+1)}}\right). \tag{6.28}$$

In a similar fashion, observing that $\tilde{Z}-E(\tilde{Z}|\tilde{X})\perp_{\mu_{X,Z}}\tilde{X}$, we obtain now

$$\begin{aligned}||r-\hat{r}||_v^2 &= E_v[(\beta-\hat{\beta})^T\tilde{X}+(\hat{f}-f)^T E(\tilde{Z}|\tilde{X})]^2\\ &\quad + E_v[(\hat{f}-f)^T(\tilde{Z}-E(\tilde{Z}|\tilde{X}))]^2,\end{aligned}$$

hence

$$E_v[(\hat{f}-f)^T(\tilde{Z}-E(\tilde{Z}|\tilde{X}))]^2 = O_P\left(\frac{\log n}{n^{2\alpha/(2\alpha+1)}}\right).$$

Define now $V=(v_{ij})_{p\times p}$, with

$$\begin{aligned}v_{ij} &= \mathrm{Cov}(Z_i-\eta_i(X),Z_j-\eta_j(X))\\ &= \mathrm{Cov}(Z_i,Z_j)-\mathrm{Cov}(\eta_i(X),\eta_j(X)),\end{aligned} \tag{6.29}$$

for $\eta_i(x) \equiv E(Z_i|X = x)$. As before, under Condition (A5) this time, V is positive definite, so

$$E_v||\hat{f} - f||_2^2 = O_P\left(\frac{\log n}{n^{2\alpha/(2\alpha+1)}}\right)$$

which implies that

$$||\hat{f_j} - f_j||_2^2 = O_P\left(\frac{\log n}{n^{2\alpha/(2\alpha+1)}}\right)$$

for $j = 1, \ldots, p$. □

References

[1] P.K. Anderson, O. Borgan, R.D. Gill, N. Keiding, Statistical Models Based on Counting Processes, Springer, Berlin, 1993.
[2] A. Barron, L. Birgé, P. Massart, Risk bounds for model selection via penalization, Probab. Theory Relat. Fields 113 (1999) 301–413.
[3] L. Birgé, P. Massart, Minimum contrast estimators on sieves: exponential bounds and rates of convergence, Bernoulli 4 (1998) 329–375.
[4] L. Birgé, P. Massart, Gaussian model selection, J. Eur. Math. Soc. 3 (2001) 203–268.
[5] F. Bunea, Penalty choices and consistent covariate selection in semiparametric regression, Florida State University, Technical Report, http://stat.fsu.edu/recentreports.html, 2002.
[6] R. De Vore, G.G. Lorentz, Constructive Approximation, Springer, Berlin, 1993.
[7] J. Fan, L. Runze, Variable selection for Cox's proportional hazards model and frailty model, Ann. Statist. 30 (2002) 74–99.
[8] D. Faraggi, R. Simon, Bayesian variable selection for censored survival data, Biometrics 54 (1998) 1475–1485.
[9] T.R. Fleming, D.P. Harrington, Counting Processes and Survival Analysis, Wiley, New York, 1991.
[10] J. Jacod, A.N. Shiryayev, Limit Theorems for Stochastic Processes, Springer, New York, 1987.
[11] F. Letué, Modéle de Cox: estimaton par sélection de modéle et modéle de chocs bivarié, Ph.D. Thesis, Laboratoire de Mathématiques de l'Université Paris-Sud, http://www.cmi.univ-mrs.fr/~letue/, 2000.
[12] T. Martinussen, T.H. Scheike, I.M. Skovgaard, Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models, Scand. J. Statist. 29 (2002) 57–74.
[13] I.W. McKeague, P.D. Sasieni, A partly parametric additive risk model, Biometrika 81 (1994) 510–514.
[14] A.E. Raftery, D. Madigan, C.T. Volinsky, Accounting for model uncertainty in survival analysis improves predictive performance, in: Bernardo et al. (Eds.), Bayesian Statistics 5—Proceedings of the Fifth Valencia International Meeting, Valencia, Spain, Oxford Science Publications, Oxford, 1996, pp. 323–349.
[15] A. Raftery, D. Madigan, C.T. Volinsky, R.A. Kronmal, Bayesian model averaging in proportional hazard models: predicting the risk of a stroke, Appl. Statist. 46 (1997) 443–448.
[16] R. Senoussi, Probléme d'identification dans le modéle de Cox, Ann. Inst. Henri Poincaré 16 (1990) 45–64.
[17] R. Tibshirani, The lasso method for variable selection in the Cox model, Statist. Med. 16 (1997) 385–395.